AI Safety Collides with the Overattribution Bias

Bruce Rushing

March 2024

Abstract

The field of Artificial Intelligence (AI) safety evaluations aims to test AI behavior for problematic capabilities like deception. However, some scientists have cautioned against the use of behavior to infer general cognitive abilities because of the human tendency to overattribute cognition to everything. They recommend the adoption of a heuristic to avoid these errors that states behavior provides no evidence for cognitive capabilities unless there is some theoretical feature present to justify that inference. We make that heuristic precise in terms of our credences's conditional independencies between behavior, cognitive capabilities, and the presence or absence of theoretical features. When made precise, the heuristic entails absurdly that failure at a behavioral task supports the presence of a theoretical feature. This is due to the heuristic suggesting inductive dependencies that conflict with our best causal models about cognition. Weakening this heuristic to allow only weak evidence between behavior and cognitive abilities leads to similar problems. Consequently, we suggest abandoning the heuristic and updating those causal models in light of the behavior observed when testing AIs for troublesome cognitive abilities.

Body Word Count: 8750
Total Word Count: 13400

1 Introduction

Artificial Intelligence (AI) safety has emerged as an important sub-discipline in computer science due to the increasing capabilities and risks associated with AIs like ChatGPT. This field aims to develop methods and safeguards against current and possible harms from AIs like the autonomous, AI-driven development of novel biological weapons. One safeguard in current use is to evaluate the threat posed by a particular AI by testing for behavior that demonstrates unwanted capabilities like autonomous replication or deception; AIs that show these undesirable traits would then be blocked from widespread deployment in commercial or government products. The sub-field of AI evaluations aims to develop better behavioral tests that can flag unwanted capabilities in new AIs. The hope is that these evaluations will eventually be adopted as standard practice in industry and government, potentially assuming the role car crash standards have in automotive transportation.

However, the method of relying on AI behavioral tests to infer worrisome cognitive abilities faces a serious problem driven by the human tendency to overattribute significant cognitive abilities to everything. Some scientists caution this overattribution bias makes empirically evaluating model behavior methodologically fraught because we are likely to see capabilities in AIs that are just not there. Consequently, they recommend an inductive heuristic that holds behavioral tests to provide no evidence for cognitive abilities unless we believe some theoretical feature or attribute is present that would be relevant to an AI possessing those cognitive abilities. This overattribution heuristic would then shield us against committing the kind of anthropomorphization errors humans are prone to.

We argue that the overattribution heuristic is not a good methodological principle for evaluating AI cognitive abilities because when it is made precise, it entails we adopt credences that conflict with our causal models about how the world works. Those causal models display our beliefs that special theoretical features are rare in the world and therefore hallmarks of true cognitive ability; these attributes are only found in certain organisms with the right sort of behavior because they are in part productive of that behavior. But this heuristic would have us believe that the absence of behavior is a hallmark of our special theoretical attributes. This is due to it recommending our credences diverge from how we think the causes work in the production of behavior. Even if we revise the overattribution heuristic to recommend that behavior provides very little evidence for cognitive hypotheses, we run into a dilemma where we commit the same error as the unrevised heuristic or we think sophisticated behavior is very common in the world. So we should avoid using the overattribution heuristic.

Here is how our argument proceeds. First, we motivate the argument by discussing problems with the sub-field of evaluation within AI safety. This leads us to describe the overattribution heuristic and why defenders of the heuristic think it is needed in any AI field that evaluates AI behavior. Second, we make exact the overattribution heuristic in terms of the probabilities they recommend to a person. Third, we specify exactly the error with the overattribution heuristic in terms of how it leads us to have an absurd credence function. Fourth, we examine a more reasonable revision of the overattribution heuristic and show this revision still leads to problems. And fifth, we suggest an alternative to the overattribution heuristic that focuses on the cognitive science of AI and the theories that best explain AI behavior.

2 The Overattribution Heuristic

An important program in the artificial intelligence safety community is the development of systematic tests to know when AIs can deceive, plan, and autonomously pursue goals that would be detrimental to humanity. This is the goal of the Model Evaluation and Research's (METR) (formerly Alignment Research Center's (ARC) Evals) team; since their launch in 2022, they have evaluated foundational models like OpenAI's GPT-4 and Anthropic's Claude. For example in METR's evaluation of GPT-4, they looked at behaviors that would indicate GPT-4 can autonomously replicate and acquire resources such as setting up an open-source language model on a server or using TaskRabbit to solve CAPTCHAs (OpenAI, 2023b; OpenAI, 2023a). Similar assessments were run on Anthropic's Claude (Anthropic, 2023, Evals, 2023). Crucially, these safety evaluations aim to infer capabilities like deception, planning, and agency in AIs—they are looking for models with the ability to *intentionally* deceive human interlocutors, to form robust plans with goals, and to do so in an autonomous fashion like human agents. It is not just aimed at problematic behavior but the underlying competencies often found in humans and more intelligent animals. This is subtle. What organizations want to find in models like GPT-4 or Claude are latent abilities to express certain behaviors that in the right conditions would lead to problematic outcomes out in the world; they aim for more than just setting up a server in a lab or having a monologue when reasoning about the use of a TaskRabbit worker but *to infer from those behaviors that the model would have the disposition to replicate without instructions or the disposition to form beliefs and deceive interacting humans.*[1] The hope is that these forms of evaluations will in the future be able to identify models with those malignant capabilities—if such abilities are discovered, AI labs could then take necessary precautions such as securing AI models and avoiding deployment.

Importantly, this strategy is fundamentally a behavioral one. Model capabilities are identified by the types of behavior the models are disposed to produce. Can a model recruit a human to overcome a security impediment? Can a model clone itself with little prompting? Can a model express to itself plans and identify key barriers to those plans' successful execution? All of these are behavioral items that teams at organizations like METR hope can be used to identify safety concerns before the model's widescale release. In essence, the strategy is no different from the accepted methods currently used to evaluate general AI capabilities; large-scale benchmarks like the Massive Multitask Language Understanding (MMLU) are used to gauge the state of the art while more human-tailored exams like the SAT and the Uniform Bar are used to infer general capabilities compared to humans (Hendrycks et al., 2021). Furthermore, anecdotal tests like the drawing of unicorns in Latex are often used to argue for significant cognitive abilities (Bubeck et al., 2023). The inference is that text-based interactions with AIs like large language models (LLMs) can show not only important cognitive abilities but also worrisome and problematic ca-

[1] This is precisely the worry that Mitchell argues METR fails to document in the infamous CAPTCHA example. METR claims GPT-4 was able to autonomously deceive a TaskRabbit worker for solving a CAPTCHA; however, Mitchell points out that both the autonomy and deception claims are weak. GPT-4 had to be vigorously prompted to do anything, with its hand held along the way, and it is unclear it even has the ability to form doxastic internal states capable of lying. She concludes that the lack of details and apparent rigorous testing methodology makes it hard to infer anything from the red team evaluations performed by METR (Mitchell, 2024).

pacities to deceive, replicate, and gain resources without human command. In short, models that act in certain ways are understood to have broader abilities found in humans and intelligent animals.

This strategy of using model behavior to make conclusions about either worrisome capacities or general cognitive abilities has received several criticisms.

One important worry is that a hazardous AI will actively deceive its evaluators about its abilities until it is deployed or gains sufficient capabilities to be extremely dangerous. The point is that an AI with some situational awareness will likely know it is being evaluated for dangerous abilities. It will then act in a way to not convince its human or AI evaluators that it can do the worrisome tasks that organizations like METR are looking for (see Christiano, 2022 and Carranza et al., 2023 for this worry and how to address it). While this is certainly a concern for the behavioral methodology of organizations like METR, we will ignore it here and focus on a separate problem.

A broader concern with utilizing behavioral tests to evaluate the safety of models comes from longstanding worries about using *any* behavioral test—especially involving text as is done with LLMs—to assess AI capabilities. At root, the problem is not so much with the models but with us: people are heavily predisposed to attribute cognitive capabilities like intentions to everything they interact with. Linguists, cognitive scientists, and computer scientists have argued that because of this human predisposition to anthropomorphize, which is sometimes called the ELIZA effect after an early computer program that fooled people into thinking it was a therapist (Weizenbaum, 1966), behavioral tests are not good tools for evaluating AI capabilities, and when they are used, they must be carefully designed and approached with caution. Speaking for this crowd, Bender and Koller write that when evaluating machines behaviorally for the ability to process meaning and understanding, scientists need to be extra cautious:

> Meaning and understanding have long been seen as key to intelligence. Turing (1950) argued that a machine can be said to "think" if a human judge cannot distinguish it from a human interlocutor after having an arbitrary written conversation with each. However, humans are quick to attribute meaning and even intelligence to artificial agents, even when they know them to be artificial, as evidenced by the way people formed attachments to ELIZA (Weizenbaum, 1966; Block 1981).
>
> This means we must be extra careful in devising evaluations for machine understanding, as Searle (1980) illustrates with his Chinese Room experiment (Bender and Koller, 2020, 5187–88).

The argument is that while tests like those proposed by Turing for evaluating intelligence have prima facie plausibility, they are methodologically worrisome because people are quick to see intelligence in everything, and we know from Searle, so they claim, that behavior is insufficient for "understanding". They go on to argue that many contemporary techniques, such as benchmarks like

MMLU, are insufficient for actually evaluating whether LLM utterances have meaning or the model has understanding. Even when models perform well on tests specifically contrived to showcase cognitive ability, they will often show glaring weaknesses in other tests that a human with the appropriate cognitive ability would not demonstrate.[2] The issue of relying on behavioral tests to gauge AI cognitive ability applies broadly to more than just LLMs. Recent work has shown that superhuman Go-playing algorithms have failed to identify a fundamental feature of playing Go that makes them vulnerable to adversarial attacks; algorithms thought to have mastered the rules and features of games like Go are exposed to have learned brittle features that break under exploitable scenarios (Wang et al., 2023).[3] The difficulty here is profound. Since humans are predisposed to seeing complex cognitive abilities with little prompting, how can we evaluate AIs for cognitive abilities from behavior alone? When should we think an AI can think, plan, or deceive if we are prone to erroneously attribute those abilities?

This suggests a methodological heuristic when evaluating AIs. Since this heuristic is aimed at avoiding overattribution errors, we call it the *overattribution heuristic*:

> **Overattribution heuristic**: Behavioral tests provide no evidence for general cognitive abilities in AIs unless there is a theoretical reason to suppose those cognitive abilities are present or absent in the AI.

By theoretical reason, we mean that the AI has a theoretical feature or attribute that suggests it has a hypothesized cognitive capacity.

The above heuristic is in operation in many arguments against relying on behavioral tests for inferring cognitive abilities. Bender and Koller argue that the lack of symbol grounding—a connection between the symbols used in language and their referents in the world—is fundamentally why we should be intensely skeptical of our predilections to attribute "meaning" and "understanding" to LLMs.[4] This same reason is appealed to by Bender et al when arguing for why LLMs should be thought of as "stochastic parrots", statistical generators

[2] Bender and Koller argue that for this reason a Turing-style test will be insufficient: "We argue that, independently of whether passing the Turing test would mean a system is intelligent, a system that is trained only on form would fail a sufficiently sensitive test, because it lacks the ability to connect its utterances to the world" (Bender and Koller, 2020, 5188).

[3] Move 37 in game two of Lee Sedol matches against AlphaGo caused the Go world-champion and commentators to attribute a genius and profound insight to AlphaGo (Metz, 2023).

[4] They write, using Searle's Chinese Room thought experiment as an illustration, that symbol grounding is lacking in LLMs:

> But language is used for communication about the speakers' actual (physical, social, and mental) world, and so the reasoning behind producing meaningful responses must connect the meanings of perceived inputs to information about that world. This in turn means that for a human or a machine to learn a language, they must solve what Harnad (1990) calls the symbol grounding problem. Harnad encapsulates this by pointing to the impossibility for a non-speaker of Chinese to learn the meanings of Chinese words from Chinese dictionary defini-

of pastiches of previous linguistic content, instead of agents with "communicative intent" familiar with how language is used, and so we should think our attributions of a genuine linguistic ability to the human predisposition to attribute meaning to any interlocutor (Bender, Gebru, et al., 2021).[5] Similar worries about inferring from model behavior sophisticated cognitive abilities infect Marcus's critique of modern LLMs and deep learning AIs generally (G. Marcus, 2018). He argues that the neural networks that underpin algorithms like GPT-4 show poor ability at transfer learning despite demonstrating competence at tasks like Atari video games; instead, a carefully controlled test often reveals the AI has only learned superficial solutions to the problems presented—solutions whose superficiality consist in failure to acquire the true rule that mastery of the

> tions alone (Bender and Koller, 2020, 5188).

The inability of people to learn a language from a dictionary alone is taken to be strong evidence that symbol grounding must be present for human understanding and so also present for an AI to understand a language.

[5] The problem here is that the type of symbol grounding humans encounter is always facilitated by interactions with a human interlocutor with genuine experiences, which helps give our words meaning, but also biases us to the ELIZA effect:

> Text generated by an LM [language model] is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. It can't have been, because the training data never included sharing thoughts with a listener, nor does the machine have the ability to do that. This can seem counter-intuitive given the increasingly fluent qualities of automatically generated text, but we have to account for the fact that our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do (Bender, Gebru, et al., 2021, 616).

However communicative intent is understood, the claim is that it must be grounded in the pragmatics of using language to interact with another interlocutor whose linguistic utterances have reference to the real world through their experience. The slogan then is that communicative intent cannot be had without symbol grounding and symbol grounding only comes through the activity of communicating with other agents whose utterances are previously grounded.

task would seem to require.[6] This is because these neural networks suffer from a fundamental inability to truly generalize (often called the out-of-distribution problem): they perform well in circumstances close to their training set but fail once something truly novel comes around (G. Marcus, 2018, 16–17). This inability to generalize Marcus attributes to contemporary connectionist architectures' failure to operate over the type of representations necessary for true generalization (see G. F. Marcus, 2003). Since neural networks lack the requisite type of representation needed for actual generalization on out-of-distribution tasks, any claim of a significant cognitive ability from behavior should be viewed with suspicion as the effect of the natural human tendency to see intelligence everywhere, i.e. an overattribution. In all three arguments, critics worried about overattribution errors apply the same heuristic of avoiding the use of behavioral data to attribute cognitive ability to certain machine learning models unless some theory suggests the behavior is informative.

Overattribution worries provide clear challenges to the use of behavioral tests for judging the safety of AI models. In the cases cited above, the authors are worried about detecting significant cognitive abilities like language comprehension; similarly, AI safety researchers want to identify dangerous cognitive abilities such as deception. But unlike the first concern of deceptive models, the challenge from scientists worried about overattribution is not that an unsafe model might be missed; instead, the problem is that a model that poses no existential risk at all because it is fundamentally dumb is inadvertently flagged by a poor testing methodology. For defenders of the overattribution heuristic, testers like METR are chasing phantasms with their behavioral methodology, which could either potentially impede the development of beneficial tools (Constantin, 2023) or ignore real harms that models like LLMs might produce (Bender, Gebru, et

[6] He gives the example of one of the early successful reinforcement learning algorithms produced by Deepmind failing to learn concepts like "ball" or "wall":

> Ostensibly, the results [of deep reinforcement learners] are fantastic: the system meets or beats human experts on a large sample of games using a single set of "hyperparameters" [....] But it is easy to wildly overinterpret what the results show. To take one example, according to a widely-circulated video of the system learning to play the brick-breaking Atari game Breakout, "after 240 minutes of training, [the system] realizes that digging a tunnel through the wall is the most effective technique to beat the game".
>
> But the system has learned no such thing; it doesn't really understand what a tunnel, or what a wall is; it has just learned specific contingencies for particular scenarios. *Transfer tests*—in which the deep reinforcement learning system is confronted with scenarios that differ in minor ways from the ones on which the system was trained show that deep reinforcement learning's solutions are often extremely superficial [....] These demonstrations make clear that it is misleading to credit deep reinforcement learning with inducing concepts like wall or paddle; rather, such remarks are what comparative (animal) psychology sometimes call overattributions (G. Marcus, 2018, 7–8).

The takeaway is that we should be careful of attributing significant mental concepts to AIs because more carefully constructed tests reveal those algorithms to have learned superficial relations and our willingness to ascribe more sophisticated abilities to these algorithms is an artifact of our psychology.

al., 2021). So the challenge then is for AI safety testers to vouchsafe their preferred methodology of behavioral testing and explain why the overattribution heuristic does not apply.

3 The Core Probabilistic Inference

Skeptics of relying on behavioral tests to assess AI capabilities appeal to the overattribution heuristic. That heuristic states that behavioral tests are not evidence for a cognitive capacity, like forming communicative intentions, having concepts, or causal reasoning, unless there is some theoretical reason to think those capacities are present. It is a heuristic to avoid the overattribution error, i.e. to avoid believing something has a power it does not. However, this claim is somewhat vague: what do we mean by " provides no evidence for" and what is the relation between "theoretical reasons" and "cognitive abilities"? We aim to precisify the claim in this section.

An important, immediate observation is that the overattribution heuristic is an inductive methodological claim. It talks about evidence and hypotheses and how the former weighs on the latter. For example, when a bullet casing is present at a crime scene, one often infers that a gun may have been fired; if we know that the type of bullet casing comes from a firearm not present in the area of the crime scene, then the casing is weak evidence for the discharge of a firearm. Like the latter inference, the heuristic merely limits when inductive evidence can be applied.

Since it is an inductive methodological claim, one can apply probability theory to make it exact. In terms of a person's probabilities, what the overattribution heuristic asks us to do is specify our posteriors or the conditional probability of the hypothesis given the evidence, $\Pr(H|E)$.

At first, one might say the heuristic is merely the claim that positive behavioral evidence should never increase our probability in a cognitive hypothesis, but negative behavioral evidence should always lower our probability in a cognitive hypothesis. This is reflected in arguments that the ability of LLMs to do certain tasks provides no evidence or worse for cognitive capacities but failure at some tasks, like transfer tests (G. Marcus, 2018, 8), is strong evidence against those cognitive capacities. In short, positive behavior does not support our hypothesis but negative behavior is strong evidence against our hypothesis.

This suggestion is problematic because it violates the principle of reflection. Reflection states that our expected posterior on some evidence should just equal our prior probability, and we follow this principle whenever we take our future probabilities to be given by our conditional probabilities.[7] The upshot is that we should never expect some evidence and its complement to both support our prior hypothesis; we are not Dr. Pangloss who holds whatever happens

[7] Reflection states our current degree of belief should be the expectation of our future degree of belief. What this means is that we expect the evidence to push us one way or another, with the push by the evidence in one direction to balance out in the other. See Huttegger, 2013 for an extended discussion.

to support our hypothesis that we live in the best of all possible worlds. We can understand our current explication of the overattribution heuristic as recommending just this fallacy: positive behavioral evidence fails to impugn the complement of our cognitive hypothesis while negative behavioral evidence suggests that the complement was correct all along. On this broken methodological recommendation, we should always believe AIs lack the requisite cognitive skills come what may. Given how silly this view would amount to, we argue that it is likely not what is meant by defenders of the overattribution heuristic.

A more charitable and philosophically interesting interpretation of the overattribution heuristic can be given in terms of conditional independence. That heuristic amounts to the twin claims about the independence of behavioral evidence and hypotheses: 1) hypotheses about cognitive abilities are independent of behavioral evidence, but 2) when conditioning on a theoretical feature or attribute behavioral evidence can strengthen or weaken the hypothesis. We go through each in turn.

Our first proposal for making the overattribution exact is that conditional on no further relevant propositions, hypotheses about cognitive abilities in AIs are probabilistically independent of behavioral evidence. Let H_c be the hypothesis about a cognitive capacity like language understanding, and let E_b be some behavioral evidence like the ability to answer questions with coherent text. Then the first part of the overattribution heuristic, "behavioral tests provide no evidence for general cognitive abilities in AIs" amounts to the independence of H_c and E_b:

$$\Pr(H_c|E_b) = \Pr(H_c) \tag{1}$$

Learning E_b without any other relevant information does not change one's credence in H_c. The worry is that without further information about the cognitive abilities of the AI, we are likely to overattribute some capacity that will not be present in the AI; to avoid that worry, the overattribution heuristic says behavioral evidence is irrelevant for the truth of a hypothesis about some sophisticated cognitive ability. It is a propaedeutic to the wild anthropomorphization bias found in humans.

This part of the heuristic comes up again and again in worries over assessing AIs from behavioral tests.

Bender and Koller, Bender et al, and Marcus all appeal to this first part of the overattribution heuristic in the course of their arguments. Bender and Koller cite Searle's Chinese room thought experiment as showing that behavioral evidence is insufficient for language understanding, and our beliefs that the Chinese room understands Chinese is due in part to our overattribution bias.[8] Similarly, they argue with a fanciful Octopus thought experiment that

[8] They write that Searle's famous thought experiment shows how we must be careful about naively attributing sophisticated mental abilities like understanding Chinese to computers:

> This [the human bias to anthropomorphization] means we must be extra careful in devising evaluations for machine understanding, as Searle (1980) elaborates with his Chinese Room experiment: he develops the metaphor of a "system"

a hyperintelligent Octopus that spies and intervenes in a conversation between two people unbeknownst to them would be able to pass most behavioral tests, but this is no indication of language understanding because, like the Chinese room, the Octopus has no access to its words' meaning due to a lack of symbol grounding.[9] Bender et al utilize the same reasoning when they declare that any behavioral evidence from LLMs for language understanding is an illusion generated by our bias to attribute communicative intent to any linguistic utterances; they argue that the coherence of text from LLMs is not evidence for the presence of any significant cognitive capacity, but an artifact of the human bias to see meaning in any text.[10] Marcus too appeals to worries about overattibution when arguing that deep learning systems have little ability to grasp concepts germane to humans. With deep reinforcement learning or language models, a model's success at beating a game or completing a linguistic task is no evidence for the model having any sophisticated understanding of the game or language because people are quick to attribute cognitive states due to the overattribu-

> in which a person who does not speak Chinese answers Chinese questions by consulting a library of Chinese books according to predefined rules. From the outside, the system seems like it "understands" Chinese, although in reality no actual understanding happens anywhere inside the system (Bender and Koller, 2020, 5188).

[9] Bender and Koller write that the hypothetical Octopus, O, can learn to approximate the daily interactions of the two people, A and B, from the words that are exchanged:

> The extent to which O can fool A depends on the task—that is, on what A is trying to talk about. A and B have spent a lot of time exchanging trivial notes about their daily lives to make the long island evenings more enjoyable. It seems possible that O would be able to produce new sentences of the kind B used to produce; essentially acting as a chatbot. This is because the utterances in such conversations have a primarily social function, and do not need to be grounded in the particulars of the interlocutors' actual physical situation nor anything else specific about the real world. It is sufficient to produce text that is internally coherent (Bender and Koller, 2020, 5188).

The issue for the Octopus comes when something new or novel is required; in the literature, this would be considered something out-of-distribution. They give a fanciful example of a coconut catapult and an encounter with a bear that the Octopus had never before seen in its textual interactions, which would push its deception abilities to the limit.

[10] They write that LLMs have no language understanding and any appearance of that is the result of human biases presenting an illusion of competency:

> Text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. It can't have been, because the training data never included sharing thoughts with a listener, nor does the machines have the ability to do that. This can seem counter-intuitive given the increasingly fluent qualities of automatically generated text, but we have to account for the fact that our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do [....] The problem is, if one side of the communication does not have meaning, then the comprehension of the implicit meaning is an illusion arising from our singular human understanding of language (independent of the model) (Bender, Gebru, et al., 2021, 616).

tion bias.[11][12] In summary, each of these arguments amounts to declaring, in probabilistic terms, that the behavioral evidence is independent of the cognitive hypothesis without conditioning on anything else of relevance.

It should be noted that if we stopped here, the overattribution heuristic would only be a negative rule in the sense it tells one when one cannot use a type of evidence for a type of hypothesis. But critics of behavioral tests are also keen to provide a recommendation for when behavioral tests should be informative; they do not want to say that behavioral evidence is *always* irrelevant. For example, Bender and Koller argue that theoretically informed behavioral tests can update towards or away from a true linguistic competency *conditional* on the fact the AI has its symbols grounded, and Marcus too would be the first to acknowledge the relevance of behavioral evidence if we know the representations of an AI can support the kind of hierarchical, symbolic structures he thinks necessary for true cognition (see G. F. Marcus, 2003 for the original idea and G. Marcus, 2020 for his most recent thoughts). So we include the caveat that behavioral evidence is informative if we have good reason to think it so.

This caveat when made exact in terms of probability theory amounts to the claim that conditional on some theoretical attribute or feature possessed by an AI, behavioral evidence can shift one's probability about a hypothesized cognitive capacity. This is just to claim that when one believes certain hypotheses about a cognitive feature or attribute of an AI, one's hypothesized cognitive capacity is now dependent on the behavioral evidence. If A_t is the proposition that the AI has a specific cognitive feature like symbol grounding and H_c and E_b are as before, then the second part of the overattribution heuristic "unless there is a theoretical reason to suppose those cognitive abilities are present in the AI" amounts to the conditional dependence of H_c on E_b given A_t:

$$\Pr(H_c|E_b, A_t) \neq \Pr(H_c|A_t) \tag{2}$$

[11] He writes that overattribution bias is the source of any claim that deep reinforcement learning agents possess concepts and that behavioral success is no evidence for true cognitive ability:

> But the system has learned no such thing; it doesn't really understand what a tunnel, or what a wall is; it has just learned specific contingencies for particular scenarios. *Transfer tests*—in which the deep reinforcement learning system is confronted with scenarios that differ in minor ways from the ones on which the system was trained to show that deep reinforcement learning's solutions are often extremely superficial [....] These demonstrations make clear that it is misleading to credit deep reinforcement learning with inducing concepts like wall or paddle; rather, such remarks are what comparative (animal) psychology sometimes call overattributions. It's not that the Atari system genuinely learned a concept of wall that was robust but rather the system superficially approximated breaking through walls within a narrow set of highly trained circumstances (G. Marcus, 2018, 8).

[12] Marcus makes an exception here for certain special behavioral tests he calls transfer tests. These do provide evidence, but they only provide evidence through our prior knowledge about the lack of neural networks processing information that symbolic representations. We discuss this below along with similar exceptions in Bender and Koller.

Learning about E_b is now informative about H_c. More specifically, the three cases frequently deployed in arguments against AI cognitive abilities are:

1. $\Pr(H_c|E_b, A_t) > \Pr(H_c|A_t)$
2. $\Pr(H_c|\neg E_b, \neg A_t) < \Pr(H_c|\neg A_t)$
3. $\Pr(H_c|E_b, A_t) > \Pr(H_c|E_b)$

Here the behavioral evidence E_b might be success answering a query or success doing a transfer test, while $\neg E_b$ is failure at answering a query or failure at a transfer test. The first case claims that positive evidence increases the probability of the hypothesis relative to the probability of the hypothesis before the evidence when we have some good theoretical reason or attribute present; the second case claims that negative evidence decreases the probability of the hypothesis relative to the probability of the evidence before the evidence when we know there is no theoretical reason or attribute present; and the third case claims that the presence of the theoretical attribute always augments the behavioral evidence. For example, case one says that we should increase our probability that an AI understands Java code conditional on the knowledge that it can compile and run Java programs if we happen to observe it matching the right outputs to inputs for a given Java program. An example of case two would be decreasing our credence that a Go-playing algorithm understands Go because we know that it lacks symbolic, recursive internal representations and we observe its failure to defeat an easily detectable strategy such as encirclement. Lastly, an example of case three would be increasing our credence that a reinforcement learning algorithm knows what a tunnel is based on its success at tunneling a video game when we discover it has the right kind of representation for variable binding. All cases illustrate that the hypotheses about AI cognitive capacities and some behavioral evidence are probabilistically dependent conditional on some theoretical reason or attribute of the AI.

Like the first part of the overattribution heuristic, the second part is often used in attacks on the naive use of behavioral evidence to underscore how behavioral evidence should inform hypotheses about cognitive abilities.

The three arguments made by Bender and Koller, Bender et al, and Marcus all rely upon the cases of the second part of the overattribution heuristic. Bender and Koller make an explicit appeal to the importance of a top-down, theoretically guided approach in natural language processing (NLP) for evaluating the successes and failures of intelligent language systems.[13] To avoid directionless hill-climbing, they recommend that NLP researchers use systematic, unified

[13] They write that much of NLP history is a repeat of the same basic process of hill-climbing that fails to succeed at producing true language understanding or general language intelligence:

> There is no doubt that NLP is currently in the process of rapid hill-climbing. [....] Thus, everything is going great when we take the bottom-up view. But from a top-down perspective, the question is whether the hill we are climbing so rapidly is the *right* hill.
>
> [....]
>
> It is instructive to look at the past to appreciate this question. [....] Researchers

theories for guiding the construction and evaluation of behavioral tests. Two examples they provide to better evaluate the behavior of AIs include the successful execution of novel Java code without a compiler or answering questions about photos; both examples provide evidence for cognitive hypotheses like language understanding because they presuppose the AI has its symbols grounded (Bender and Koller, 2020, 5189–5190). Their examples illustrate case one, two and three of the second part of the overattribution heuristic by showing that the cognitive hypothesis of language understanding can only be supported or defeated by behavioral evidence through either the presence or absence of symbol grounding in the AI and that presence of symbol grounding always improves the support provided by behavioral evidence. Bender et al make a similar move by dismissing the apparent success of LLMs on benchmarks because LLMs are only trained on data involving signs and have no access to the meaning of those signs; the implicit argument is that the lack of meaning indicates that certain tests are telling against LLM capabilities and if meaning were present somehow in the training data, then we should infer cognitive hypotheses like language understanding.[14] This is just all three cases where the truth of a theoretical proposition like the dataset has a link between linguistic form and meaning determines when behavioral evidence strengthens or weakens the probability of the cognitive hypothesis, i.e. language understanding, and always does so. Marcus makes the same argument too by appealing to the importance of the existence of symbolic representations in AIs for utilizing behavioral evidence as evidence towards hypotheses about cognitive ability. While a model's successful passing of a behavioral test provides no evidence of a significant cognitive capacity, the failure on special transfer tests is evidence because it leverages a model's lack of recursively structured representations to enable true generalization performance on out-of-distribution tasks. We know the behavioral test is relevant because we know deep learning algorithms cannot compute the special type of symbolic representations.[15] This is just the application of these three cases: it is case

> of each generation felt they were solving relevant problems and making constant progress, from a bottom-up perspective. However, eventually serious shortcomings of each paradigm emerged, which could not be tackled satisfactorily with the methods of the day, and these methods were seen as obsolete. This negative judgment—we were climbing a hill, but not the right hill—can only be made from a top-down perspective (Bender and Koller, 2020, 5191).

[14] Bender et al write authoritatively that no linguistic understanding occurs in LLMs and tie the failure of LLMs on sensitive tests to the lack of a connection between the form and meaning of a sign:

> However, no actual language understanding is taking place in LM-driven approaches to these tasks, as can be shown by careful manipulation the test data to remove spurious cues the systems are leveraging [21, 93]. Furthermore, as Bender and Koller [14] argue from a theoretical perspective, languages are systems of signs [37], i.e. pairings of form and meaning. But the training data for LMs is only form; they do not have access to meaning. Therefore, claims about model abilities must be carefully characterized (Bender, Gebru, et al., 2021, 615).

[15] Marcus writes that this is ultimately the reason why we should be skeptical of any claim

one because an AI with the right sort of symbolic representation would make successful behavioral tests relevant to the hypothesis, it is case two because AIs without the right sort of symbolic representation like neural networks can be demonstrated to be cheats through carefully designed transfer tests, and it is case three because success at a transfer test is always stronger evidence when the theoretical attribute is present as opposed to being absent. So all three authors rely upon the second part of the overattribution heuristic to tell us when behavioral evidence happens to be relevant to cognitive hypotheses.

In summary, the overattribution heuristic can be made precise through two claims about the probabilities concerning cognitive hypotheses H_c, behavioral evidence E_b, and some theoretical feature or attribute A_t:

1. $\Pr(H_c|E_b) = \Pr(H_c)$
2. $\Pr(H_c|E_b, A_t) \neq \Pr(H_c|A_t)$
    (a) $\Pr(H_c|E_b, A_t) > \Pr(H_c|A_t)$
    (b) $\Pr(H_c|\neg E_b, \neg A_t) < \Pr(H_c|\neg A_t)$
    (c) $\Pr(H_c|E_b, A_t) > \Pr(H_c|E_b)$

The first states the behavioral evidence is no evidence concerning the cognitive hypothesis and the second states it becomes evidence when we condition on some theoretical feature or attribute. The second claim is cashed out more specifically in terms of how we think the evidence should move us with regards to the cognitive hypothesis; when we have reason to believe the theoretical feature or attribute is present in an AI, then positive behavioral evidence should always strengthen our belief in the cognitive hypothesis, but when we think that feature or attribute is absent, then negative behavioral evidence should weaken our belief in the cognitive hypothesis. That is just to say that we avoid the overattribution bias by declining to let behavioral tests inform us about cognitive capacities in AIs unless we have good reason to think those behavioral tests are relevant.

4 The Core Problem

The overattribution heuristic recommends we treat behavioral evidence as probabilistically independent of our cognitive hypotheses unless we condition on the

that a neural network has a true cognitive feature:

> The core problem, at least at present, is that deep learning learns correlations between sets of features that are themselves "flat" or nonhierarchical, as if in a simple, unstructured list, with every feature on equal footing. Hierarchical structure (e.g., syntactic trees that distinguish between main clauses and embedded clauses in a sentence) are not inherently or directly represented in such systems, and as a result deep learning systems are forced to use a variety of proxies that are ultimately inadequate, such as the sequential position of a word presented in a [sic] sequences (G. Marcus, 2018, 10).

presence or absence of some special theoretical feature. The "unless" part more precisely says that the presence of a theoretical attribute and some positive behavioral evidence should always raise our credence in the cognitive hypothesis, and it says that the absence of a theoretical attribute and some negative behavioral evidence should lower our credence in the cognitive hypothesis. These parts of the heuristic are often combined with the following inference relating behavior and those theoretical attributes. Would-be defenders of the overattribution heuristic also say positive and negative behavior should also be evidence for and against those special theoretical features that make that evidence relevant for cognitive hypotheses. Intuitively, we should think that a model coherently forming grammatical sentences is some evidence—however weak—for it having symbol grounding or employing symbolic representations, while weakness at specific behaviors is also evidence—perhaps stronger—for it lacking those same features. We don't think rocks have symbolic representations in part because they fail to perform any sophisticated behavior; but chimpanzees might have those representations due to their sophisticated problem-solving and communicative skills. So we have good reason to think that our beliefs in the theoretical attributes we think relevant to learning about cognition are responsive in some degree to the behaviors we observe.

However, the overattribution heuristic prevents us from taking the intuitive relationship between behavioral evidence and theoretical attributes. Its four parts of probabilistic independence between hypotheses and evidence plus the specific direction of the conditional probabilities for evidence and the attribute entail that we should think negative behavioral evidence provides higher credence to the presence of theoretical attributes than positive behavioral evidence:

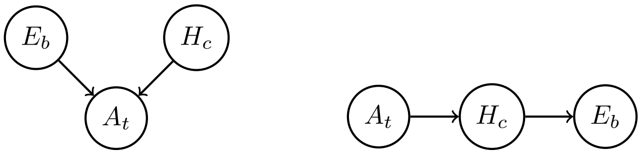
Proposition 1. *If Pr obeys the following four properties of the overattribution heuristic:*

1. $Pr(H_c|E_b) = Pr(H_c)$
2. $Pr(H_c|E_b, A_t) > Pr(H_c|A_t)$
3. $Pr(H_c|\neg E_b, \neg A_t) < Pr(H_c|\neg A_t)$
4. $Pr(H_c|E_b, A_t) > Pr(H_c|E_b)$

Then $Pr(A_t|\neg E_b) > Pr(A_t|E_b)$.[16]

What proposition 1 tells us is that if we adopt the overattribution heuristic, then *we must hold negative behavioral evidence to be more of a sign for theoretical attributes than positive behavioral evidence.* But this is absurd. It means that we should think it more likely that a trash can has symbol grounding or symbolic representation after it fails to answer our questions; conversely, an adult undergraduate who does answer our questions is less likely than we had

[16] We would like to thank Anonymous for pointing out and initially proving this result.



(a) Collider

(b) Chain

Figure 1: The graphs that the overattribution heuristic and our causal model give us. The graph in (a) is induced by the heuristic, while the graph in (b) is what our causal theory suggests.

previously thought to have symbol grounding or symbolic representations. Behavioral evidence still tells us something about the theoretical attributes: it just tells us in a direction incompatible with how we think the world works.

This points to a deeper problem with the overattribution heuristic. If we picture our propositions and their probabilistic relations graphically, we find they form a *collider* (see figure 1a). A collider is a directed acyclic graph where two parent variables point towards a common sink variable. This graphical structure is induced by the conditional independencies of our probability function. For example, if our two parent variables are the outcomes of two fair coin tosses and the sink variable is the ringing of a bell whenever one of the coins lands head, then we will find that while our probabilities over the outcomes of the coin tosses are independent, they will be connected once we condition on the bell; if we hear the bell and find one coin to have landed tails, then we know that the other coin must have landed heads. The problem is that this graphical structure highlights a system of independencies given by the overattribution heuristic that our best causal model of the world would imply is false. Our best causal picture has our propositions form a *chain* (see figure 1b). We think the theoretical attribute is somehow productive of behavior and whether it is mediated by our cognitive hypothesis or not, this entails a certain dependence in the causal model that our credence function as given by the overattribution heuristic just doesn't have. Hence why we find it weird to think that negative behavioral evidence should be a reason to believe the presence of a theoretical attribute.

It should be emphasized what is at issue is not a disagreement over the causes. Instead, we have a tension between how we think some of our probabilities should behave given the causes and what the overattribution heuristic tells us should be the case. The probabilities between those are different—even though we do not think the collider is our causal structure. This is subtle. The overattribution heuristic does not tell us that the correct causal model is a collider; rather, the collider is a reflection of our credences independence structure as given by the heuristic. However, if we believe as do the proponents of the overattribution heuristic that the correct causal model is given by the chain, then our probabilities should be otherwise when relating behavior and theoretical features. So even though we are not contradicting ourselves about the causal structure, we end up believing absurd things about the relationship between behavioral evidence and the theoretical features of cognition that are absurd *qua* our causal models.

A natural suggestion then would be to keep the second part of the overattribution heuristic while abandoning one of the specific claims relating the movement of our credences in the hypotheses given the theoretical attributes and evidence. There are three options. Either we abandon condition two, condition three, or condition four in proposition one. All options are unpalatable to defenders of the overattribution heuristic. The first would entail that our credences in a cognitive hypothesis should go down conditional on the presence of a theoretical attribute and some positive evidence.[17] When we see a human with symbol grounding utter coherent sentences, then we should think it less probable that they have linguistic understanding; or when we see an AI with a sophisticated symbolic representation play Atari games, this would make us think it lacks true concepts relative to not observing it play Atari games. So this is no good. What about option two? Changing this condition would entail that knowing an AI lacks a theoretical feature and observing negative behavioral evidence should increase our credence in the truth of the cognitive hypothesis. When a table fails to answer our linguistic queries and we have good reason to think it lacks symbol grounding, well then we should update towards it having linguistic competence! Similarly, a neural network that we know lacks symbolic representation and fails at a game of Go should push us towards believing it in fact knows the concepts of Go. Like before, this is untenable for defenders of the overattribution heuristic. Finally, we can abandon condition four. But abandoning this would be to argue that a theoretical feature we think crucial to cognition, like symbol grounding, *can be absent and provide more support to a cognitive hypothesis than otherwise.* So there is some behavioral evidence out there such that the absence of symbol grounding makes it more plausible that an AI has true linguistic understanding; witnessing a mechanical Turk playing chess—when we know there is a person underneath it—can lead us to conclude that the fraudalent mechanism knows a thing or two about chess. But this is absurd. Consequently, abandoning any of these options will not work.

That leaves only one option: abandon the independence of the behavioral evidence from the cognitive hypothesis. This would mean that observing behavior should tell us about the presence or absence of a cognitive ability like language understanding or manipulation of concepts. But we take this to be a wholesale abandonment of the overattribution heuristic; after all, the heuristic protects us from overattribution bias by making our credences in cognitive hypotheses impervious to behavioral evidence. So our recommendation is to jettison the overattribution heuristic generally and learn about cognitive hypotheses differently.

[17] We could not set them equal as that would entail independence—negating the whole reason for the second part of the overattribution heuristic.

5 Approximating Independence

Abandoning the independence of behavioral evidence from cognitive hypotheses equates to a rejection of the overattribution heuristic. However, what if instead of requiring independence we only allowed behavioral evidence to provide slight inductive support to cognitive hypotheses? This amounts to a less stringent overattribution heuristic; we merely caution the scientist about naively relying upon behavioral evidence while still allowing that evidence to be substantial in the presence of other known factors. The revised overattribution heuristic can then be given as:

> **Revised Overattribution heuristic**: Behavioral tests provide *very little* evidence for general cognitive abilities in AIs unless there is a theoretical reason to suppose those cognitive abilities are present or absent in the AI.

Of course, we need to be more precise by what we mean by "very little" evidence in the revised heuristic. We turn to that now.

Again, by adopting the framework of probability theory, we can make this revised heuristic precise. The latter half of the heuristic is as before—namely, we say that behavioral evidence increases the probability of a hypothesis conditional on some theoretical attribute relative to the probability of that hypothesis conditioned on that attribute alone, negative behavioral evidence decreases our probability of the cognitive hypothesis when we know the theoretical attribute is absent, and we think the presence of a theoretical attribute always enhances the truth of a cognitive hypothesis given some behavioral evidence. The new condition we add in substitution for the independence of the cognitive hypothesis and the behavioral evidence is that *the evidence gained when observing some behavior is approximately zero.* The evidence gained is simply the difference between our conditional probability of the cognitive hypothesis given the behavioral evidence and the marginal probability of the cognitive hypothesis alone, i.e. $\Pr(H_c|E_b) - \Pr(H_c)$. The claim is that this equals some epsilon, greater than zero but approximately zero:

$$\Pr(H_c|E_b) - \Pr(H_c) = \epsilon, \quad \epsilon > 0 \text{ and } \epsilon \approx 0 \tag{3}$$

This says that positive behavioral evidence is still evidence, just so weak that it hardly matters as an update on the truth of the cognitive hypothesis; observing an LLM write a poem about a peanut butter sandwich stuck in a VCR is *slight* evidence that it understands what a peanut butter sandwich is or how it would fit into a VCR, but it is negligible evidence at best. While we do not require independence between the evidence and hypothesis, we do require that for most practical purposes that evidence is irrelevant. This seemingly prevents us from committing an overattribution error because we are only allowed to weakly update on our behavioral observations unless we have good reason to take those observations seriously.

Some approaches to studying AIs' cognitive abilities already recommend this revised overattribution heuristic. It is a natural step in the approach recommended by Bender, Marcus, and others. And it has appeared recently in discussions around using AIs like LLMs as models for human cognition. For example, Pavlick, 2023 recommends that caution be applied to inferring sophisticated cognitive abilities to LLMs and to surmising those LLMs cognize in the same manner as humans. While behavioral evidence is still evidence, it is only very weak evidence unless modulated by theory and tells us little about the underlying competence of LLMs and how that competence is implemented relative to how humans achieve a similar competence:

> Until we can precisely characterize the representations and mechanisms in play under the hood, examples of LLMs' behavioural successes or failures tell us little about LLMs' ability to serve as models of language in humans. Of course, it can be argued that requiring analysis of the internal processing of LLMs amounts to holding LLMs to a higher bar than that to which we hold humans. We obviously cannot inspect humans' internal neurological processing with the level of precision or invasiveness at which we can in principle inspect LLMs. It is true that this is a higher bar, and to a large extent, that is the entire point. If we want to consider LLMs, or any computational model, as a candidate model of the human mind, we must know something about how they work under the hood. Black box predictive models do little to advance understanding. Importantly, though, this higher bar holds whether we want to make positive or negative claims. Until we understand how LLMs work, we cannot assert that their internal processing bears any resemblance to humans, but we also cannot assert that it bears no resemblance. Undoubtedly, a precise characterization of neural networks' representations and mechanisms is not trivial to acquire and will take time. Work is already happening that brings us closer to characterizing this internal structure [...], and once such findings are mature, we can reanalyse this behavioural evidence and draw much stronger conclusions, positive or negative (Pavlick, 2023, 2).

Pavlick's interest here is slightly different from Bender and Marcus's but the recommendation is the same in spirit: she cautions against the naive application of success at behavioral tasks, and while her recommendation for treating that behavioral evidence is not quite as strong as Bender and Marcus, she still adopts their suggestion that evidence matters more when the underlying mechanisms are known. Black boxing models and relying on their success at certain behavioral tasks provides some justification for the cognitive hypothesis that LLMs manipulate representations like how humans manipulate representations but that justification is very little until we have a better model of the mechanisms in LLMs and their similarity to human mechanisms. Once those mechanisms—those theoretical features—have been understood the behavioral evidence can better support or refute hypotheses about LLM cognition.

While this suggestion has the immediate appeal that it does not directly lead to the absurdity of the original overattribution heuristic, it leads to a troublesome dilemma. Supposing that the evidence provided by some behavior is positive towards a hypothesis, the following equality holds:

$$(\mathrm{Pr}(H_c|E_b) - \mathrm{Pr}(H_c))\mathrm{Pr}(E_b) = (\mathrm{Pr}(H_c) - \mathrm{Pr}(H_c|\neg E_b))\mathrm{Pr}(\neg E_b)$$

Treating the difference $\mathrm{Pr}(H_c|E_b) - \mathrm{Pr}(H_c)$ as some positive, approximately zero ϵ, we can easily show[18] that:

$$\mathrm{Pr}(H_c|E_b) = \frac{\epsilon}{\mathrm{Pr}(\neg E_b)} + \mathrm{Pr}(H_c|\neg E_b) \tag{4}$$

Two facts should be observed about this equation.

First, if the ratio between ϵ and the negative behavioral is approximately zero, then it results in a near equality holding between $\mathrm{Pr}(H_c|E_b)$ and $\mathrm{Pr}(H_c|\neg E_b)$. This means that in most cases we will have the undesirable result that negative behavioral evidence provides more support for the presence of a theoretical attribute than positive behavioral evidence, i.e. $\mathrm{Pr}(A_t|\neg E_b) > \mathrm{Pr}(A_t|E_b)$, because this is exactly the condition that independence allows proposition one to hold. This can be shown with the following proposition:

Proposition 2. *If Pr obeys the following four properties:*

1. $Pr(H_c|E_b) > Pr(H_c)$
2. $Pr(H_c|E_b, A_t) > Pr(H_c|A_t)$
3. $Pr(H_c|\neg E_b, \neg A_t) < Pr(H_c|\neg A_t)$
4. $Pr(H_c|E_b, A_t) > Pr(H_c|E_b)$

[18] Applying the definition of the probability of a negation we have:

$$\begin{aligned}
\epsilon\mathrm{Pr}(E_b) &= (\mathrm{Pr}(H_c) - \mathrm{Pr}(H_c|\neg E_b)\mathrm{Pr}(\neg E_b)) \\
\epsilon\mathrm{Pr}(E_b) &= (\mathrm{Pr}(H_c|E_b)\mathrm{Pr}(E_b) + \mathrm{Pr}(H_c|\neg E_b)\mathrm{Pr}(\neg E_b) - \mathrm{Pr}(H_c|\neg E_b)\mathrm{Pr}(\neg E_b)) \\
\frac{\epsilon}{\mathrm{Pr}(\neg E_b)} &= \mathrm{Pr}(H_c|E_b) + \frac{\mathrm{Pr}(H_c|\neg E_b)\mathrm{Pr}(\neg E_b) - \mathrm{Pr}(H_c|\neg E_b)}{\mathrm{Pr}(E_b)} \\
\mathrm{Pr}(H_c|E_b) &= \frac{\epsilon}{\mathrm{Pr}(\neg E_b)} - \frac{\mathrm{Pr}(H_c|\neg E_b)\mathrm{Pr}(\neg E_b) - \mathrm{Pr}(H_c|\neg E_b)}{\mathrm{Pr}(E_b)} \\
\mathrm{Pr}(H_c|E_b) &= \frac{\epsilon}{\mathrm{Pr}(\neg E_b)} - \frac{\mathrm{Pr}(H_c|\neg E_b)(1 - \mathrm{Pr}(E_b)) - \mathrm{Pr}(H_c|\neg E_b)}{\mathrm{Pr}(E_b)} \\
\mathrm{Pr}(H_c|E_b) &= \frac{\epsilon}{\mathrm{Pr}(\neg E_b)} - \frac{\mathrm{Pr}(H_c|\neg E_b) - \mathrm{Pr}(H_c|\neg E_b)\mathrm{Pr}(E_b) - \mathrm{Pr}(H_c|\neg E_b)}{\mathrm{Pr}(E_b)} \\
\mathrm{Pr}(H_c|E_b) &= \frac{\epsilon}{\mathrm{Pr}(\neg E_b)} + \mathrm{Pr}(H_c|\neg E_b)
\end{aligned}$$

And if $Pr(A_t|\neg E_b) < Pr(A_t|E_b)$ *and* $Pr(H_c|E_b) - Pr(H_c|\neg E_b) = \delta$ *where* $\delta > 0$ *then, either:*

$$Pr(H_c|E_b, A_t) - Pr(H_c|\neg E_b, A_t) < \delta$$

or:

$$Pr(H_c|E_b, \neg A_t) - Pr(H_c|\neg E_b, \neg A_t) < \delta$$

What proposition 2 says is that if the evidence supports the presence of the theoretical attribute more than the absence of evidence, then it has to be the case that our conditional probabilities relating the cognitive hypothesis, theoretical attribute, and behavioral evidence are bound by the ratio of ϵ and $\Pr(\neg E)$: when we suppose that ratio to be small, we have to think the evidence provides little relative support to the hypothesis even when we condition on the theoretical attribute. Putting aside the question of whether this is good inferential practice, clearly most cases will not behave this way when that ratio is approximately zero. The upshot is that in the majority of cases, we will be led to the strange inference that negative behavioral evidence is evidence for theoretical features and attributes crucial to cognition. Thus when the ratio between the evidence we gain from observing some behavior and our prior probabilities on the absence of that behavior are small—as will be the case when we doubt the truth of an AI passing a behavioral test—we are led to the absurdity that failure at a task supports the presence of sophisticated cognitive mechanisms or features.

Second, if the ratio between ϵ and the negative behavioral is not approximately zero, then *the marginal probability of the behavioral evidence is very high.* This follows from the law of the negation of probabilities; the ratio of ϵ and $\Pr(\neg E_b)$ is not approximately zero when $\Pr(\neg E_b)$ is close to ϵ, i.e. close to zero, and since $\Pr(\neg E_b) = 1 - \Pr(E_b)$, $\Pr(E_b)$ must be close to one. We must think the particular behavioral evidence is very likely: answering questions, writing poems, solving math problems, and so on are common behaviors we expect to find out in the world.

The upshot is that we are faced with a dilemma if we adopt the revised overattribution heuristic. Either we think the ratio of evidence gained from observing an AI perform a certain behavior relative to the marginal probability of it not producing that behavior is sufficiently small as to be trivial or it is not. If trivial, then we end up believing that the absence of that behavior should increase our confidence in the presence of the AI possessing an important theoretical feature or attribute; if not trivial, then we must think that particular behavior is very likely to have occurred regardless of the cognitive features of that AI. The former case leads to absurdity in how we infer theoretical features are in the world while the latter makes us think sophisticated behavior is just common in the world—in avoiding overattributions of beliefs, desires, understanding, and other sophisticated cognitions to objects in the world, we make unusual and rare behavior commonplace in nature. So we should not feel that the revised overattribution heuristic provides any good guidance for combating the human tendency to anthropomorphize.

6 Discussion

The overattribution heuristic or its revised cousin should not guide the methodology for evaluating AI capabilities, including safety capabilities. But we still have the problem the heuristic was meant to remedy. Namely, how do we combat the human cognitive bias to see agency and mental life in things that just do not have them? This worry is still present before AI safety evaluators who hope to leverage behavioral tests for detecting dangerous AI capabilities.

A promising solution comes from the observation that our credence in cognitive hypotheses is the product of a mixture of the causal models we think have plausibility. In this picture, we know by the law of total probability that our credence in a cognitive hypothesis is the sum of our conditional probabilities of that hypothesis given some causal model multiplied by our prior in that model. So when we observe some new evidence, we update those likelihoods and priors by the evidence to infer how we should think about our cognitive hypotheses.

This has two methodological recommendations for combating the overattribution bias.

First, we should simply consider how relevant behavioral evidence is for a cognitive capacity according to our best models about how that behavior might be produced and how likely we think those models are true. Over time, our estimates about the importance of a particular piece of behavioral evidence can change as we change our minds about the probability of our causal models. If we keep track carefully our priors on those models and what those models say, then we can avoid an overattribution error; after all, the claim that we make overattribution errors is often driven by a belief that our best causal models do not support our attribution of significant cognitive ability. So we should not automatically discount behavioral evidence but weigh how we think that behavior would have been produced given our theories.

Secondly and importantly, this means that unexpected behavior evidence should also inform us about the correctness of our causal theories. If some bit of behavior is produced in a way that we did not expect given our priors over the causes, well then we should be good Bayesians and *change our mind about those causes*. The relevance of evidence for safety concerns is also the relevance of evidence for how we think cognition works. Sudden behavior such as deception and lying by an LLM whose cognitive architecture fails to correspond to our leading causal theories should lead us to downweight those theories in favor of alternatives that can better account for that behavior.

This means that AI safety work that proceeds by evaluating behavior is intimately tied together with the cognitive sciences. Persons working at METR and other organizations that rely upon these tests should know well what our best theories of cognition are and what we would expect given those theories; but linguists, neuroscientists, psychologists, and cognitive scientists should also be willing to change their mind about how interesting behavior is produced when that behavior occurs in things that seemingly do not reflect their favorite theories.[19] Researchers in the AI safety space can perform valuable scientific work while also being confident that they can learn about model abilities from behavior.

The recommendation we have for addressing the challenge of anthropomorphization to behavioral methodology in AI safety is to have a renewed focus on what our theories of cognition say and how they can be wrong. Bias can only be defeated by following sound inductive methodology and not appealing to heuristics that conflict with that methodology and our best theories about the world.

Appendix 1

To prove Proposition 1, recall the four parts of the overattribution heuristic:

1. $\Pr(H_c|E_b) = \Pr(H_c)$
2. $\Pr(H_c|A_t, E_b) > \Pr(H_c|A_t)$
3. $\Pr(H_c|\neg A_t, \neg E_b) < \Pr(H_c|\neg A_t)$
4. $\Pr(H_c|E_b, A_t) > \Pr(H_c|E_b)$

From symmetry of independence, these three parts amount to the following conditions as well:

5. $\Pr(H_c|\neg E_b) = \Pr(H_c)$
6. $\Pr(H_c|A_t, E_b) > \Pr(H_c|A_t, \neg E_b)$
7. $\Pr(H_c|\neg A_t, E_b) > \Pr(H_c|\neg A_t, \neg E_b)$
8. $\Pr(H_c|E_b, A_t) > \Pr(H_c|E_b, \neg A_t)$

We can then apply the law of total probability to $\Pr(H_c|E_b)$:

$$\begin{aligned}\Pr(H_c|E_b) &= \Pr(H_c, A_t|E_b) + \Pr(H_c, \neg A_t|E_b)\\ &= \Pr(H_c|A_t, E_b)\Pr(A_t|E_b) + \Pr(H_c|\neg A_t, E_b)\Pr(\neg A_t|E_b)\end{aligned}$$

[19] Mitchell similarly recommends AI researchers collaborate with cognitive scientists to design better tests and evaluate LLMs and other AI models to counteract overattribution bias (Mitchell, 2023). However, she fails to recommend that cognitive scientists should change their minds about their current theories based on AI performance. This is a difference between our recommendation here and the one offered by her: she is way more confident in the explanations and models cognitive scientists have about how reasoning operates. Constructing better behavioral tests is a good recommendation, but the measure of "better" here is a measure conditional on a theory of what are the cognitive mechanisms that produce that behavior. Testing systematic variations of the same task works to support a cognitive hypothesis only if we think that something like abstract variable binding is productive of that cognitive hypothesis. It may turn out that variable binding is neither necessary nor sufficient for an intelligent agent—and may not even be present in humans.

And the law of total probability to $\Pr(H_c|\neg E_b)$:

$$\begin{aligned}\Pr(H_c|E_b) &= \Pr(H_c, A_t|\neg E_b) + \Pr(H_c, \neg A_t|\neg E_b)\\ &= \Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b) + \Pr(H_c|\neg A_t, \neg E_b)\Pr(\neg A_t|\neg E_b)\end{aligned}$$

Since $\Pr(H_c|E_b) = \Pr(H_c) = \Pr(H_c|E_b)$ we have:

$$\begin{aligned}&\Pr(H_c|A_t, E_b)\Pr(A_t|E_b) + \Pr(H_c|\neg A_t, E_b)\Pr(\neg A_t|E_b) =\\ &\quad \Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b) + \Pr(H_c|\neg A_t, \neg E_b)\Pr(\neg A_t|\neg E_b)\end{aligned} \tag{5}$$

We claim that conditions 6 and 7 plus this equality imply that $\Pr(A_t|E_b) < \Pr(A_t|\neg E_b)$. Suppose not for contradiction. Then a) $\Pr(A_t|E_b) = \Pr(A_t|\neg E_b)$ or b) $\Pr(A_t|E_b) > \Pr(A_t|\neg E_b)$.

For a), note that condition 6 and a) implies:

$$\begin{aligned}&\Pr(H_c|A_t, E_b) > \Pr(H_c|A_t, \neg E_b)\\ &\Pr(H_c|A_t, E_b)\Pr(A_t|E_b) > \Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b)\\ &\Pr(H_c|A_t, E_b)\Pr(A_t|E_b) + \Pr(H_c|\neg A_t, E_b)\Pr(\neg A_t|E_b) >\\ &\quad \Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b) + \Pr(H_c|\neg A_t, E_b)\Pr(\neg A_t|E_b)\end{aligned}$$

And condition 7 and a) implies:

$$\begin{aligned}&\Pr(H_c|\neg A_t, E_b) > \Pr(H_c|\neg A_t, \neg E_b)\\ &\Pr(H_c|\neg A_t, E_b)\Pr(\neg A_t|E_b) > \Pr(H_c|\neg A_t, \neg E_b)\Pr(\neg A_t|E_b)\\ &\Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b) + \Pr(H_c|\neg A_t, E_b)\Pr(\neg A_t|E_b) >\\ &\quad \Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b) + \Pr(H_c|\neg A_t, \neg E_b)\Pr(\neg A_t|E_b)\\ &\Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b) + \Pr(H_c|\neg A_t, E_b)\Pr(\neg A_t|E_b) >\\ &\quad \Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b) + \Pr(H_c|\neg A_t, \neg E_b)\Pr(\neg A_t|\neg E_b)\end{aligned}$$

But this results in the following inequality:

$$\begin{aligned}&\Pr(H_c|A_t, E_b)\Pr(A_t|E_b) + \Pr(H_c|\neg A_t, E_b)\Pr(\neg A_t|E_b) >\\ &\quad \Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b) + \Pr(H_c|\neg A_t, E_b)\Pr(\neg A_t|E_b) >\\ &\qquad \Pr(H_c|A_t, \neg E_b)\Pr(A_t|\neg E_b) + \Pr(H_c|\neg A_t, \neg E_b)\Pr(\neg A_t|\neg E_b)\end{aligned} \tag{6}$$

Equation 6 contradicts equation 5 since the l. h. s. and the r. h. s. are supposed to be equal.

For b), we assume $\Pr(A_t|E_b) > \Pr(A_t|\neg E_b)$:

$\Pr(H_c|A_t,E_b)\Pr(A_t|E_b)+\Pr(H_c|\neg A_t,E_b)\Pr(\neg A_t|E_b)$
$=\Pr(H_c|A_t,E_b)\Pr(A_t|E_b)+\Pr(H_c|\neg A_t,E_b)(1-\Pr(A_t|E_b))$
$=\Pr(H_c|A_t,E_b)\Pr(A_t|E_b)+\Pr(H_c|\neg A_t,E_b)-\Pr(H_c|\neg A_t,E_b)\Pr(A_t|E_b))$
$>\Pr(H_c|A_t,E_b)\Pr(A_t|E_b)+\Pr(H_c|\neg A_t,E_b)-\Pr(H_c|\neg A_t,E_b)\Pr(A_t|\neg E_b))$ $(\star)$
$>\Pr(H_c|A_t,E_b)\Pr(A_t|E_b)+\Pr(H_c|\neg A_t,\neg E_b)-\Pr(H_c|\neg A_t,\neg E_b)\Pr(A_t|\neg E_b))$ $(\dagger)$
$>\Pr(H_c|A_t,\neg E_b)\Pr(A_t|\neg E_b)+\Pr(H_c|\neg A_t,\neg E_b)-\Pr(H_c|\neg A_t,\neg E_b)\Pr(A_t|\neg E_b))$ $(\ddagger)$
$=\Pr(H_c|A_t,\neg E_b)\Pr(A_t|\neg E_b)+\Pr(H_c|\neg A_t,\neg E_b)(1-\Pr(A_t|\neg E_b))$
$=\Pr(H_c|A_t,\neg E_b)\Pr(A_t|\neg E_b)+\Pr(H_c|\neg A_t,\neg E_b)\Pr(\neg A_t|\neg E_b)$

Step $\star$ follows from our assumption, step $\dagger$ follows from condition 7, and step $\ddagger$ follows from our assumption and condition 8. The result contradicts equation 5. □

Appendix 2

To prove Proposition 2, recall the four parts conditions we should have satisfy:

1. $\Pr(H_c|E_b) > \Pr(H_c)$
2. $\Pr(H_c|E_b, A_t) > \Pr(H_c|A_t)$
3. $\Pr(H_c|\neg E_b, \neg A_t) < \Pr(H_c|\neg A_t)$
4. $\Pr(H_c|E_b, A_t) > \Pr(H_c|E_b)$

These can be restated as:

5. $\Pr(H_c|E_b) > \Pr(H_c|\neg E_b)$
6. $\Pr(H_c|A_t, E_b) > \Pr(H_c|A_t, \neg E_b)$
7. $\Pr(H_c|\neg A_t, E_b) > \Pr(H_c|\neg A_t, \neg E_b)$
8. $\Pr(H_c|E_b, A_t) > \Pr(H_c|E_b, \neg A_t)$

Suppose that $\Pr(A_t|\neg E_b) < \Pr(A_t|E_b)$ and $\Pr(H_c|E_b) - \Pr(H_c|\neg E_b) = \delta$ where $\delta > 0$. We need to show that either $\Pr(H_c|A_t, E_b) - \Pr(H_c|A_t, \neg E_b) < \delta$ or $\Pr(H_c|\neg A_t, E_b) - \Pr(H_c|\neg A_t, \neg E_b) < \delta$. We aim to show this that it cannot be the case that those two differences are both greater than δ.

To see build an intuition for why this is the case, consider the diagrams in figures 2 and 3. Here we see $\Pr(H_c|E_b)$ and $\Pr(H_c|\neg E_b)$ are mixtures, as given by the lines between α ($\Pr(H_c|A_t, E_b)$) and β ($\Pr(H_c|\neg A_t, E_b)$) and γ ($\Pr(H_c|A_t, \neg E_b)$) and λ ($\Pr(H_c|\neg A_t, \neg E_b)$). The relative ordering is fixed by those Greek letters, which correspond to the conditional probabilities of

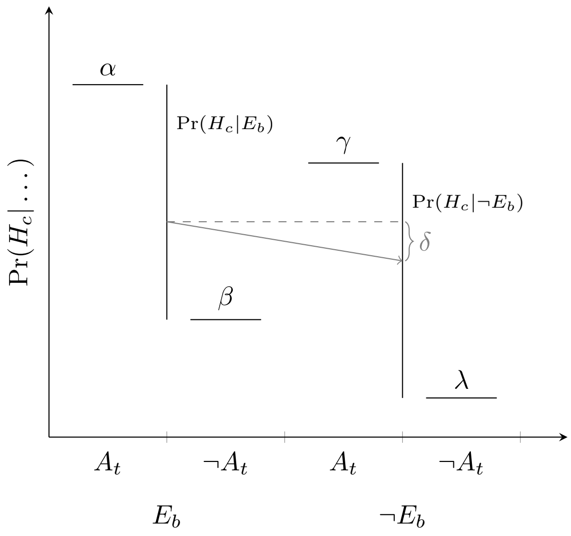


Figure 2: Two mixtures of conditional probabilities $\mathrm{Pr}(H_c|E_b)$ and $\mathrm{Pr}(H_c|\neg E_b)$. The y-axis gives the conditional probability in H_c in the given proposition. The x-axis corresponds to where in the algebra the proposition is true. The legend for the values are: $\alpha = \mathrm{Pr}(H_c|A_t, E_b)$, $\beta = \mathrm{Pr}(H_c|\neg A_t, E_b)$, $\gamma = \mathrm{Pr}(H_c|A_t, \neg E_b)$, and $\lambda = \mathrm{Pr}(H_c|\neg A_t, \neg E_b)$. δ is as in the below proof, which here must be positive. The arrowed line between the two mixtures represents their difference.

$\mathrm{Pr}(H_c| \ldots$ for the different combinations of A_t and E_b. The value δ indicates the difference between the two mixtures when $\mathrm{Pr}(H_c|E_b) > \mathrm{Pr}(H_c)$. We have a latitude then for the value of $\mathrm{Pr}(A_t|E_b$ to be greater than $\mathrm{Pr}(A_t|\neg E_b)$ indicated by the arrow going from the first mixture to the second; naturally, this will only occur when δ is big enough. In figure 2 and 3, that δ will only be big enough when it is either greater than $\alpha - \gamma$ or it is greater than $\beta - \lambda$.

To prove this proposition, we consider first the case where $\gamma > \lambda$ and then the other case when $\lambda \geq \gamma$.[20] Note that we can compute the values of $\mathrm{Pr}(H_c|E_b)$ and $\mathrm{Pr}(H_c|\neg E_b)$ by either decreasing α and γ by some value respectively or increasing β and λ by some value respectively. This results in the following functions:

$$\mathrm{Pr}(H_c|E_b) = \alpha - a \tag{7}$$

$$\mathrm{Pr}(H_c|\neg E_b) = \gamma - b \tag{8}$$

$$\mathrm{Pr}(H_c|E_b) = \beta + c \tag{9}$$

[20] We would like to thank Anonymous for helping with this proof.

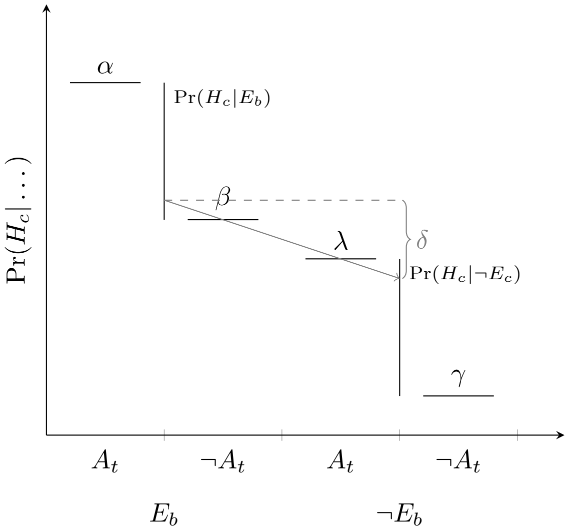


Figure 3: Two mixtures of conditional probabilities $\Pr(H_c|E_b)$ and $\Pr(H_c|\neg E_b)$. The y-axis gives the conditional probability in H_c in the given proposition. The x-axis corresponds to where in the algebra the proposition is true. The legend for the values are: $\alpha = \Pr(H_c|A_t, E_b)$, $\beta = \Pr(H_c|\neg A_t, E_b)$, $\gamma = \Pr(H_c|A_t, \neg E_b)$, and $\lambda = \Pr(H_c|\neg A_t, \neg E_b)$. δ is as in the below proof, which here must be positive.

$$\Pr(H_c|\neg E_b) = \lambda + d \tag{10}$$

This results in δ being expressed by the following equations:

$$\delta = \alpha - a - \gamma + b \tag{11}$$

$$\delta = \beta + c - \lambda - d \tag{12}$$

We can then bound the differences by rearranging equations 11 and 12:

$$\alpha - \gamma = \delta + a - b \tag{13}$$

$$\beta - \lambda = \delta + d - c \tag{14}$$

This means that $\alpha - \gamma < \delta$ just when $a < b$ and similarly $\beta - \lambda < \delta$ just when $d < c$. We aim to show it cannot be the case that both are greater than or equal to δ. To show this, we need to show that it cannot be the case that $a \geq b$ and $d \geq c$.

To show that, suppose for contradiction that $a \geq b$ and $d \geq c$. Then note that we can actually compute a, b, c, d from the law of total probability. For example, a can be found:

$$\begin{aligned}\Pr(H_c|E_b) &= \alpha(1-x)+\beta x\\ &= \alpha - \alpha x + \beta x\\ &= \alpha - (\alpha x - \beta x)\\ &= \alpha - (\alpha - \beta)x\end{aligned}$$

where $x = \Pr(\neg A_t|E_b)$. Similarly, we find that all of the constants are:

$$a = (\alpha - \beta)\Pr(\neg A_t|E_b) \tag{15}$$

$$b = (\gamma - \delta)\Pr(\neg A_t|\neg E_b) \tag{16}$$

$$c = (\alpha - \beta)\Pr(A_t|E_b) \tag{17}$$

$$d = (\gamma - \delta)\Pr(A_t|\neg E_b) \tag{18}$$

Now consider the ratios between a and c and b and d:

$$\begin{aligned}\frac{a}{c} &= \frac{(\alpha-\beta)\Pr(\neg A_t|E_b)}{(\alpha-\beta)\Pr(A_t|E_b)}\\ &= \frac{\Pr(\neg A_t|E_b)}{\Pr(A_t|E_b)}\end{aligned}$$

$$\begin{aligned}\frac{b}{d} &= \frac{(\gamma-\delta)\Pr(\neg A_t|\neg E_b)}{(\gamma-\delta)\Pr(A_t|\neg E_b)}\\ &= \frac{\Pr(\neg A_t|\neg E_b)}{\Pr(A_t|\neg E_b)}\end{aligned}$$

Note, from our assumption that $\Pr(A_t|E_b) > \Pr(A_t|\neg E_b)$ and the negation rule for probabilities, it follows that

$$\frac{a}{c} < \frac{b}{d} \tag{19}$$

$$ad < bc \tag{20}$$

But our assumptions imply that $ad \geq bc$ since a, b, c, d are all positive and $a \geq b$ and $d \geq c$. Contradiction.

For the other case of the proof where $\gamma \leq \lambda$, we use the same reasoning except now we need to change the ordering of our $\Pr(H_c|\neg E_b)$ functions:

$$\Pr(H_c|\neg E_b) = \lambda - b \tag{21}$$

$$\Pr(H_c|\neg E_b) = \gamma + d \tag{22}$$

This changes our equalities with respect to δ:

$$\delta = \alpha - a - \gamma - b \tag{23}$$

$$\delta = \beta + c - \lambda + d \tag{24}$$

Rearranging equations 23 and 24 we have:

$$\alpha - \gamma = \delta + a + b \tag{25}$$

$$\beta - \lambda = \delta - c - d \tag{26}$$

Importantly, equation 26 implies that $\beta - \lambda \leq \delta$ in all cases. We should merely note that c and d cannot both be zero since they equal:

$$c = (\alpha - \beta)\Pr(A_t|E_b)$$

$$d = (\delta - \gamma)\Pr(A_t|\neg E_b)$$

and since we assume $\alpha > \beta$ and $\delta > \gamma$ and $\Pr(A_c|E_b) > \Pr(A_c|\neg E_b)$ implies that one of these quantities is strictly positive. So it will always be the case that $\beta - \lambda < \delta$ and by disjunction introduction, we trivially complete the proof. □

References

Anthropic (2023). *Model Card and Evaluations for Claude Models.* https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

Bender, Emily M, Timnit Gebru, et al. (2021). “On the dangers of stochastic parrots: Can language models be too big?” In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.

Bender, Emily M and Alexander Koller (2020). “Climbing towards NLU: On meaning, form, and understanding in the age of data”. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198.

Bubeck, Sébastien et al. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4.* arXiv: 2303.12712 [cs.CL].

Carranza, Andres et al. (2023). “Deceptive Alignment Monitoring”. In: *The Second Workshop on New Frontiers in Adversarial Machine Learning.* URL: https://openreview.net/forum?id=obsO44GFhh.

Christiano, Paul (2022). *Mechanistic anomaly detection and ELK.* URL: https://www.alignmentforum.org/posts/vwt3wKXWaCvqZyF74/mechanistic-anomaly-detection-and-elk (visited on 09/20/2023).

Constantin, Sarah (2023). *Why I am Not An AI Doomer.* URL: https://sarahconstantin.substack.com/p/why-i-am-not-an-ai-doomer (visited on 09/26/2023).

Evals, ARC (2023). *Update on ARC's recent eval efforts.* URL: https://evals.alignment.org/blog/2023-08-01-new-report/ (visited on 09/20/2023).

Hendrycks, Dan et al. (2021). "Measuring Massive Multitask Language Understanding". In: *International Conference on Learning Representations.* URL: https://openreview.net/forum?id=d7KBjmI3GmQ.

Huttegger, Simon M (2013). "In defense of reflection". In: *Philosophy of Science* 80.3, pp. 413–433.

Marcus, Gary (2018). "Deep learning: A critical appraisal". In: *arXiv preprint arXiv:1801.00631.*

— (2020). "The next decade in AI: four steps towards robust artificial intelligence". In: *arXiv preprint arXiv:2002.06177.*

Marcus, Gary F (2003). *The algebraic mind: Integrating connectionism and cognitive science.* MIT press.

Metz, Cade (2023). *In Two Moves, AlphaGo and Lee Sedol Redefined the Future.* URL: https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/ (visited on 09/26/2023).

Mitchell, Melanie (2023). "How do we know how smart AI systems are?" In: *Science* 381.6654, eadj5957. DOI: 10.1126/science.adj5957. eprint: https://www.science.org/doi/pdf/10.1126/science.adj5957. URL: https://www.science.org/doi/abs/10.1126/science.adj5957.

— (2024). *Did GPT-4 Hire and Then Lie To a Task Rabbit Worker to Solve a CAPTCHA?* URL: https://aiguide.substack.com/p/did-gpt-4-hire-and-then-lie-to-a (visited on 01/07/2024).

OpenAI (2023a). "GPT-4 System Card". In: pp. 40–99. arXiv: 2303.08774 [cs.CL].

— (2023b). *GPT-4 Technical Report.* arXiv: 2303.08774 [cs.CL].

Pavlick, Ellie (2023). "Symbols and grounding in large language models". In: *Philosophical Transactions of the Royal Society A* 381.2251, p. 20220041.

Wang, Tony Tong et al. (2023). "Adversarial Policies Beat Superhuman Go AIs". In.

Weizenbaum, Joseph (1966). "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 9.1, pp. 36–45.